星魂黄页网爬虫目录如下: yellowpageSpider-for-github\yellowpage



1.

 定义抓取的网页信息条目，例如：楼盘名，网址等… 用类字典表示

2

 爬虫scrapy中个性化设置

3.

 数据库定义及写入的管道 pipelines.py连接sqlite数据库

pipeline2.py连接本地mysql数据库

4.

 星魂网一级、二级、三级目录对应的页面网址汇总

5.  定义爬虫的抓取具体逻辑实现

6.  定义爬虫中间件：

其中，

rotate\_proxy为代理ip中间件

rotate\_useragent为代理user-agent中间件

7.  存放user-agent池和ip池

其中，crawl-proxy-nonblock.py为获取ip代理的脚本文件;

useragentlist.txt为UA池

validProxy为代理ip池

8.

 爬取到的数据

9. 数据库连接文件

